The Double Marking of Liberal Studies in the Hong Kong Public Examination System

David CONIAM
The Chinese University of Hong Kong

Abstract

Background: This article reports a study into the double marking of Liberal Studies in Hong Kong. This is now a compulsory subject in Hong Kong's Years 10-12 curriculum which, when first examined in the new Hong Kong Diploma of Secondary Education in 2012, will increase its candidature from its current 3,300 to 80,000.

Aims: To examine the reliability of the forthcoming double marking of LS, investigating whether high inter-marker correlations are achieved in the double marking of LS, matching the reliability rate achieved by other examinations administered by the HKEAA.

Method: Broadly adopting the methodology of an earlier study of marking in the 2007 Year 11 English Language examination, the current study investigates double marking using classical test statistics – inter-, intra- and marker-subject correlations – and the amount of discrepancy between pairs of markers.

Sample: Seven experienced markers (re)marked 677 scripts from the *Hong Kong Studies* module and another seven markers (re)marked 654 scripts from the Human Relationships module.

Results: Moderate to strong correlations emerged between pairs of markers. Discrepancy levels were below 10% – in line with other Hong Kong public examinations.

Conclusion: With a view to improving marker reliability, the study concludes with the recommendation that the current holistic marking scheme should be re-evaluated, with a view to investigating its replacement by an analytic, domain-based marking scheme. This study highlights the need for public examination bodies to carry out a range of validation, reliability and other studies prior to the implementation of changes to their large-scale examinations.

Keywords: Double marking, liberal studies, inter-marker correlations

香港公開考試制度：通識教育科雙閱卷員制

龔大胃
香港中文大學

摘要

*背景：*本文報告一項有關香港通識教育科雙閱卷員制 (double marking)的研究。通識教育現在是一個在香港高中一年級至三年級課程的必修科目，它會在第一屆的香港中學文憑試（即2012年）內首次評核，屆時其考生人數會從目前的3300人增至80000人。

*目的：*查證即將實施的通識教育科雙閱卷員制的可靠性，以調查其閱卷員之間能否取得高相關性，以及其可靠性能否相匹配其他由香港考試評核局管理的考試。

*方法：*廣泛採用的方法是利用較早前在2007年有關高中二年級英國語文考試的研究所採用的，目前的研究使用標準的測驗統計 - 閱卷員及科目的間信度和內信度，以及每對閱卷員之間的差異水平之大小。

*樣本：*七個富經驗的閱卷員（重新）評核677份來自單元〈香港研究〉的考卷，另七個富經驗的閱卷員（重新）評核654份來自單元〈人際關係〉的考卷。

*結果：*閱卷員之間取得中度到高度的相關性。閱卷員之間的差異水平低於 10% - 與其他香港公開考試一致。

*結論：*為了提高閱卷員的可靠性，這個研究結論建議將目前整體性的評分準則重新評估，以便調查能否以分析性、範圍性的評分準則取代之。這項研究結果強調了公開考試機構在實施應用在大規模考試的改變前是有需要進行一系列的有關有效性、可靠性和其他方面的研究。

關鍵詞：雙閱卷員制、通識教育、閱卷員之間的相關性

Introduction

In Hong Kong, major changes to the educational and examination systems came into effect in September 2009, when the Hong Kong secondary school system changed from a 5+2 model (i.e., seven years, as in the British system) to a 3+3 model (i.e., six years, similar to the Chinese and Australian education models). Further, rather than students sitting public examinations in Years 11 and 13, from 2012 onwards there will be a single end-of-high-school examination at the end of Year 12 – the Hong Kong Diploma of Secondary Education (HKDSE), the annual candidature for which will be approximately 80,000. To compensate for the loss of one year in secondary school education, tertiary education will increase from three years to four years.

Accompanying the structural changes to the senior secondary curriculum, there have been qualitative changes to the curriculum itself. One of these involves the introduction of Liberal Studies (LS) as a compulsory subject. The introduction of LS – a subject which aims to promote and develop students' 'critical thinking' in the understanding of social issues (Ip, 2010) – has aroused a substantial amount of controversy in Hong Kong (Chan, 2005; Tsang, 2006), not least of which is the discussion surrounding how such a 'critical thinking', 'subjective in nature', 'textbook independent' subject will be marked (Kuo, 2007). Consequently, any issue related to how LS is examined, or indeed marked, is viewed with great interest by stakeholders. See Coniam & Yeung (2010) for an elaboration of the position of LS in the Hong Kong senior secondary curriculum.

A number of major changes are being implemented along with the structural changes accompanying the new Year 10-12 curriculum and examinations. One of these is the adoption of onscreen marking for all subjects from 2010 onwards (see Coniam, 2009a, 2010). Another is the adoption of double marking for the compulsory major subjects – LS, Chinese Language and Culture, and English (for which double marking has long been an established feature). Since the focus of the current study involves double marking, the following section examines the issue of the double marking of examination scripts.

Double Marking

The debate over the marking of essays has a long history. Over fifty years ago, Pilliner (1969) referred to the 'notorious unreliability of essay marking' (p. 313). While it has been argued in some places that double marking does not increase the reliability of marking (Cox, 1967), the general consensus takes the opposite view, stating that double marking does provide for an increase in reliability, with researchers demonstrating this position statistically, e.g., Pilliner (1969), Brooks and Linton (2004).

In this context, studies over the past decade have investigated whether double marking is intrinsically more reliable than single marking. Some double marking studies indicate that more subjective forms of assessment report lower marker agreement. Newton (1996), for example, investigating Mathematics and English General Certificate of Secondary Education (GCSE) essay-based examinations in the UK, reported much lower reliability for English than for Mathematics. More recently, Vidal Rodeiro (2007) describes a GCSE study involving the double marking of papers in two different subjects requiring extended responses: *Verse Literature* in Classical Greek, and *Literary Heritage and Imaginative Writing* in English Language. In the Classical Greek paper, markers were more restricted in that they had to award marks for answers specified in the marking scheme; in contrast, in the English paper, markers had to evaluate the quality of candidates' free responses (ibid, p. 33). Higher reliability emerged from

the Classical Greek marking than from the English examination marking. As a coda to this issue, Linton (2004) presents a statistical illustration of how a 'truer' score for a candidate may be arrived at through double marking.

Brooks (2004) presents a cogent update of the whys and hows of double marking. She makes two important points. The first is that, over the past two decades in the UK, double marking has all but vanished in the marking of public examination scripts – despite being taken up in the tertiary sector (e.g., Partington, 1994). The second is that double marking also appears to have suffered from a dearth of research activity.

Brooks offers two reasons for the 'decline in interest in double marking' (ibid, p. 31). The first concerns the supply of examiners. Citing a 10-fold increase in the amount of GCSE O and A level scripts over the past three decades, she states that there are simply not sufficient examiners for the numbers of scripts that need to be marked. The second reason is closely allied to the first, and concerns the costs incurred through having to pay two markers rather than one. Brook concludes that operationally, therefore, in monitoring the validity and reliability of public examinations in the UK "… double marking offers, perhaps, the least likely way forward." (ibid, p. 42). It would appear that this approach is underpinned by logistical rather than statistical reasons.

Brooks calls for fresh research in the area of double marking, with a view to providing information as to whether double marking serves a valid purpose. She cites the double marking of Mathematics, for example, echoing Newton's (1996) research indicating that double marking in this subject would achieve very little in terms of improved reliability.

The current study therefore positions itself as furthering agendas raised in Brooks' paper concerning double marking. In Hong Kong, the across-the-board introduction of the subject LS has raised a considerable amount of concern (Rong, 2005; Lam & Zhang, 2005). This has been due, first, to the size of the cohort and, second, to discussions of how the subjective construct of analytical and critical thinking, in the context of extended responses, will produce fair and reliable assessments of candidates. To this end, the current study extends the research agenda into double marking in the field of LS in the Hong Kong context. Double marking has long been a feature of the marking of the major languages in Hong Kong public examinations, having been introduced in the Year 11 Hong Kong Certificate of Education Examination (HKCEE) in 1980 (King, 1980; Coniam, 1991). All English language Writing examinations are now double marked – at Year 11 HKCEE and Year 13 Hong Kong Advanced Level Examination (HKALE). Chinese language Writing is double marked at HKCEE, and as from 2012 in the Year 12 HKDSE, the Chinese language Practical Writing examination will be double marked.

The major focus of analysis in the current study is on reliability as measured through inter-rater correlations. There are other ways of examining the reliability of marking – Rasch being one such statistic (see Coniam, 2009b). The current study utilises inter-rater (Pearson Product Moment) correlations since the study focuses on a subject in the Hong Kong public examination system whose principal method of assessing reliability in performance assessments in the Hong Kong context is monitored by the Hong Kong Examinations and Assessment Authority through inter-rater correlations as well as correlations with other papers (Choi & Lee, 2010; King, 1994). Another way in which reliability may be gauged is through the number of scripts that may need to be remarked if there is a large discrepancy between two markers on any given script.

As described below, the HKEAA's standard practice is to remark scripts (i.e., call for a third marker) where the inter-marker discrepancy is approximately greater than 20% of the maximum possible mark (5/25 marks for LS). Consequently, a description of mark discrepancy in LS also forms part of the analysis in the current study.

Hypotheses

There are two related hypotheses in the current study. The first hypothesis, related to the reliability of the marking, is that there will be high inter-marker correlations – comparable to those achieved in other examinations administered by the Hong Kong Examinations and Assessment Authority (HKEAA). While interpretations of correlation magnitude differ, the current study will follow Hatch and Lazaraton's definitions (1991: 441) for inter-rater reliability whereby a "strong" correlation is taken as 0.8 or above, a "moderate to strong" correlation 0.5-0.8, and a "moderate" correlation 0.5. The first hypothesis is supplemented by a second hypothesis: that discrepancies between the two markers will be less than 10% of the total scripts marked (comparable to the situation existing in the public English language Writing examinations).

The Study

The data used in the study are drawn from the 2009 HKALE LS examination, where the candidature was 3,307. While there are six modules in the LS examination, the current project – to contain the scope of the study – focuses on two of the largest marking panels: Hong Kong Studies and Human Relationships [Note 1].

Seven experienced markers from each of these two panels, who had marked in the (single-marked) 2009 HKALE LS examination in May, were invited to take part in the double-marking exercise in December 2009, eight months later. The methodology adopted in the current study focused on the three compulsory questions of the two modules. Markers would each re-mark, on paper, approximately 100 of the compulsory questions from the 2009 examination, with each batch containing a number of scripts markers had marked previously. This procedure has been used successfully before (Coniam, 1991; Coniam 2009); the time lag of eight months is sufficiently long for markers not to recollect having marked the scripts before, rendering them as unfamiliar as unseen scripts.

The significance of the study is that after the diminution, world-wide, of support for double marking, the topic is being re-visited in the light of a) issues of reliability; b) a massive growth in candidature; c) the cost benefit implications of such practice in terms of the expense and effort that it requires; d) the implications of the discrepancies that arise in double marking processes for comparable examinations; and e) its implications for other public examination authorities.

Data Analysis

The methodology for analysis broadly mirrors the classical test theory (CTT) approach reported in Coniam (2009a). Statistics used – following standard HKEAA practice – were inter- and intra-marker correlations together with correlations with an external paper serving as an objective 'anchor'. While the major focus of the examination centres around critical thinking, there is a language element involved – the *Effective Communication* marking subscale. This, in a sense, is comparable to one of the reliability measures conducted by the HKEAA for English which involves correlating the scores of the subjectively marked *Writing* paper with the objectively-marked (i.e., limited response) Reading paper (Choi & Lee, 2010, p. 68). Although LS has no objectively-marked paper, the *Effective Communication*

score is language-related; consequently, the scores for LS are correlated against the scores on the Chinese Language and Culture (CLC) examination.

Available reference points for double-marked performance assessments in the Hong Kong public examination system are as follows. The inter-marker correlation for the Year 13 2009 HKALE Use of English *Writing* examination (candidature 36,000, 115 markers) was 0.75. The inter-marker correlation for the Year 11 HKCEE English language *Writing* paper in 2006 (candidature 77,000, 188 markers) was 0.79. The inter-marker correlation for the 2007 HKCEE Chinese language *Writing* paper (candidature 82,000, 336 markers) for 2007 was 0.69. The inter-marker correlations obtained by the HKEAA for its English language examinations have generally been in the 0.7-0.8 range (Choi & Lee, 2010, p. 69). The correlations for English presented above are consistent with this, approximating to the desired level (Hatch and Lazaraton, 1991) of 0.8. The correlation figures for Chinese fall below the levels achieved for English, being more moderate than strong.

The inter-paper correlation between the 2009 HKALE Use of English *Writing* paper and the objectively-marked *Reading and Language Systems* paper was 0.74.

Concerning discrepancy rates, for the 2007 HKCEE English language *Writing* paper, where 5 points out of the 24 available was the trigger for the third marker, approximately 10% of scripts required a third marking (see Coniam, 2009a). In the 30-marker study of onscreen versus paper-based marking (see Coniam, 2009a), an overall discrepancy marking figure of 8.1% emerged.

Markers, Candidates and Modules

Of the 14 participating markers, 3 had been marking LS for one year or less, 6 for 2-5 years and 3 over 6 years. Concerning age, 2 were below 30, 11 were in the 31-50 range, and 1 was over 50. There were 8 males and 6 females. All had been teaching LS for at least 3 years, and 11 were their school's HKALE LS coordinator. These figures broadly reflect the larger population of the 49 markers who marked the 2009 LS examination (see Coniam & Yeung, 2010).

The candidate sample was selected so that it represented a cross-section of the ability range for both modules. Table 1 details the sample.

Table 1

Sample

Module	Question	Number	Topic
Hong Kong Studies	1	239	Teaching Chinese through Putonghua
	2	244	Opening up radio broadcasting
	3	194	Voters' political orientations
Total		677	
Human Relationships	1	182	Conducting friendships online
	2	190	Students participating in civic activities
	3	282	Marketing strategies of tutorial schools
Total		654	

As can be seen from Table 1, each question has a total sample of between 200-300: 677 scripts (against a target of 700) for Hong Kong Studies and 654 for Human Relationships. There is, nonetheless, some variation in the spread of candidates over the six questions of the two modules. In the two-module dataset of 1,331 scripts, there were 990 individual candidates, with 356 candidates (36.0%) sitting both Hong Kong Studies and Human Relationships modules. This picture reflects the LS examination generally, with the two modules being taken by 39.8% of the candidature in 2009 and by 40.1% in 2010.

Inter- and Intra-Marker Correlations

The modules are marked independently of one other; consequently, results for each module are presented separately. The results of inter-marker correlations conducted between markers and against Chinese language and culture (CLC) are presented in Table 2.

Table 2

Inter-Marker Correlations

		Hong Kong Studies					Human Relationships		
		Mkr 1 total	Mkr 2 total	CLC			Mkr 1 total	Mkr 2 total	CLC
Marker 1 total	r		.590	.288	Marker 1 total	r		.614	.254
	sig.		.000	.000		sig.		.000	.000
	N		677	664		N		654	639
Marker 2 total	r	.590		.219	Marker 2 total	r	.614		.344
	sig.	.000		.000		sig.	.000		.000
	N	677		664		N	654		639

Note. Mkr=Marker; CLC =Year 13 HKALE Chinese Language and Culture

Table 2 shows that the inter-marker correlations would be classified (after Hatch & Lazaraton) as moderate to strong for *Hong Kong Studies* [0.59], and *Human Relationships* [0.61]. The correlations are, however, not as high as those obtained for English and Chinese: the inter-marker correlations were 0.75 and 0.79 for English, and 0.69 for Chinese. Given that the marking schemes are similar for both English and Chinese (an analytic scheme comprising four domains), this difference in correlations is worthy of further exploration when variables such as the length, quality, depth and comparative success of marker experience, qualifications and training can be investigated and analysed.

As 5/25 marks are allocated to *Effective Communication*, and 94% of the candidature chose to answer in Chinese [Note 2], Table 2 provides, as an external reference point, the correlations between the markers' total scores and the subjects' Chinese language results. As can be seen, these correlations, although significant, are small. Chinese correlates with LS at 0.2-0.3. As noted above, the inter-marker correlation for Chinese is lower than the figure obtained for English. Since the score range available in the *Effective Communication* correlation is small – only 5 points – a lower correlation might be expected.

Tables 3a and 3b below explore correlations in more depth, investigating the inter- and intra-marker question subpart correlations, as well as the correlations between question subparts with *Effective Communication*.

Table 3a

Hong Kong Studies – Q. subpart correlations

Q1. Teaching Chinese through Putonghua						
		M1 (b)	*M1 EC*	*M2 (a)*	*M2 (b)*	*M2 EC*
M1 (a)	R	.444	.429	.375	.194	.315
	Sig.	.000	.000	.000	.003	.000
	N	239	239	239	239	239
M1 (b)	R		.413	.181	.509	.275
	Sig.		.000	.005	.000	.000
	N		239	239	239	239
M2 (a)	R				.355	.397
	Sig.				.000	.000
	N				239	239
M2 (b)	R					.465
	Sig.					.000
	N					239

Q2. Opening up radio broadcasting						
		M1 (b)	*M1 EC*	*M2 (a)*	*M2 (b)*	*M2 EC*
M1 (a)	R	.326	.478	.538	.304	.274
	Sig.	.000	.000	.000	.000	.000
	N	244	244	244	244	244
M1 (b)	R		.592	.179	.617	.446
	Sig.		.000	.005	.000	.000
	N		244	244	244	244
M2 (a)	R				.281	.570
	Sig.				.000	.000
	N				244	244
M2 (b)	R					.563
	Sig.					.000
	N					244

Q3. Voters' political orientations						
		M1 (b)	*M1 EC*	*M2 (a)*	*M2 (b)*	*M2 EC*
M1 (a)	R	.479	.312	.206	.192	.141
	Sig.	.000	.000	.004	.007	.049
	N	194	194	194	194	194
M1 (b)	R		.593	.233	.565	.493
	Sig.		.000	.001	.000	.000
	N		194	194	194	194
M2 (a)	R				.355	.439
	Sig.				.000	.000
	N				194	194
M2 (b)	R					.735
	Sig.					.000
	N					194

Table 3b

Human Relationships – Q. subpart correlations

Q1. Conducting friendships online						
		M1 (b)	*M1 EC*	*M2 (a)*	*M2 (b)*	*M2 EC*
M1 (a)	R	.427	.618	.554	.358	.407
	Sig.	.000	.000	.000	.000	.000
	N	182	182	182	182	182
M1 (b)	R		.648	.246	.669	.510
	Sig.		.000	.001	.000	.000
	N		182	182	182	182
M2 (a)	R				.462	.662
	Sig.				.000	.000
	N				182	182
M2 (b)	R					.650
	Sig.					.000
	N					182

Q2. Students participating in civic activities						
		M1 (b)	*M1 EC*	*M2 (a)*	*M2 (b)*	*M2 EC*
M1 (a)	R	.249	.605	.605	.224	.423
	Sig.	.001	.000	.000	.002	.000
	N	190	190	190	190	190
M1 (b)	R		.488	.250	.674	.491
	Sig.		.000	.000	.000	.000
	N		190	190	190	190
M2 (a)	R				.280	.611
	Sig.				.000	.000
	N				190	190
M2 (b)	R					.492
	Sig.					.000
	N					190

Q3. Marketing strategies of tutorial schools						
		M1 (b)	*M1 EC*	*M2 (a)*	*M2 (b)*	*M2 EC*
M1 (a)	R	.388	.369	.516	.216	.267
	Sig.	.000	.000	.000	.000	.000
	N	282	282	282	282	282
M1 (b)	R		.529	.301	.600	.370
	Sig.		.000	.000	.000	.000
	N		282	282	282	282
M2 (a)	R				.426	.474
	Sig.				.000	.000
	N				282	282
M2 (b)	R					.618
	Sig.					.000
	N					282

Note. M1 = Marker 1; M2 = Marker 2; EC = Effective Communication

As Tables 3a and 3b show, the inter-marker correlations for each part of each question are generally around 0.5. For *Hong Kong Studies*, the inter-marker correlations for Part (a) vary from 0.2-0.5; for *Hong Kong Studies*, Part (b), the figure is 0.5-0.6. For *Human Relationships*, the inter-marker correlations for Parts (a) and (b) is around 0.5-0.6. These figures and the discussion that follows should be treated somewhat cautiously, however. The reason is that question subparts are generally marked out of a possible 10 marks – providing a possible 20-mark total. However, there are only 5 marks available for *Effective Communication*. Such small allocations of marks may skew correlations.

The intra-marker correlations between a given marker and their marking of the two parts of a question, i.e., Part (a) with Part (b), correlate at a moderate level, generally around the 0.3-0.4 level.

The correlations between the mark for each part and *Effective Communication* show slightly different trends, with *Effective Communication* correlations generally comparable, if at times slightly higher, than the subpart correlations with each other. For *Hong Kong Studies*, most correlations for Q.1 are around 0.4, although there is a spread from 0.3-0.5 on this question. On the other questions, the correlation is around 0.5, although 0.7 is achieved on Q.3. With *Human Relationships*, the majority of the correlations are around 0.6, although on Q.3 only a 0.3 correlation is achieved. As can be seen, then, Part (b) questions generally correlate higher – often at the 0.6 level. This might be due to the fact that whereas Part (a) requires candidates to interpret a piece of information such as a diagram or a piece of text, Part (b) requires them to extrapolate and to think critically.

Currently, the marking scheme for LS has seven broad levels, specified holistically. This may well leave many of the decisions regarding score awards to individual marker interpretation – despite lengthy markers' meetings before marking gets underway. In contrast, English, has four subscales (each consisting of six levels) which more explicitly define the constructs against which markers award marks. Research has shown that inexperienced raters in particular are more able to apply criteria laid out in separate analytic scales than in holistic scales (see Weir, 1990). The higher correlation in the case of English may in part therefore be attributed to the more explicit nature of the marking domains.

LS is an integrated task – a 'performance assessment' involving reasoning and problem solving (see Harmon et al., 1997, p.5). In an analysis of how LS functions as a performance assessment in the Hong Kong public examination context, Kuo (2007) asserts that the subject should be assessed as such, i.e. with a domain-specified analytic rubric, which will be a more robust marking instrument than holistic marking.

Discrepancies Between the Two Forms of Marking

A further extension of the inter-marker reliability issue can be observed in the amount of discrepancy between marks awarded to particular scripts. Lower correlations between markers will tend to generate more discrepancies. The system in place for English is that a third marker is invoked to resolve large discrepancies between markers. The following section discusses the discrepancy marking context for LS in the current study. A common criterion for invoking re-marking (i.e., the use of a third marker) has been established as two markers differing from each other by more than one score point on a six-point scale (see e.g., Attali & Burstein, 2005, p. 13). As mentioned above, for the 2007 HKCEE English language *Writing* paper (with a difference of 5/24 points between the two markers being

the trigger) approximately 10% of the scripts underwent third marking.

Table 4 below presents the discrepancy rates between the two markers' scores for the two modules studied.

Table 4

Differences Between Markers' Scores

	Hong Kong Studies	Human Relationships	Total discrepancies
Level of discrepancy	75/667 (11.2%)	55/654 (8.4%)	130/1,331 (9.8%)

As can be seen from Table 4, compared with the overall discrepancy rate for the 2007 HKCEE English language *Writing* paper of 10%, and 8.1% for the English language onscreen/paper-based marking comparative study (Coniam, 2009a), the overall discrepancy figure of 9.8% for LS is comparable with the figures for English. One interesting outcome is that the discrepancy rate for *Human Relationships*, at 8.4%, is substantially lower than the figure of 11.2% for *Hong Kong Studies*. Given the moderate-to-strong 0.6 correlations between markers, it might have been assumed that discrepancy rates would be larger. A figure at the 10% level can therefore be considered acceptable. However, it should be noted that the current study has involved 14 experienced markers. In marked contrast, the enlarged candidature in 2012, will require many more markers, all of whom will be inexperienced. Whether the 10% discrepancy level can be maintained will need to be closely monitored, and rigorous training and possibly an expanded use of monitoring scripts in the marking process may be required (see Coniam, 2009a).

Conclusion

This article has described an exploratory study of the double marking of two modules in the existing HKALE Liberal Studies examination. The current study has been exploratory in nature since both the format (i.e., module structure etc) as well as the marking (single vs. double) of the HKALE and HKDSE examinations are different. Consequently, it is not practically possible to provide a direct comparison between the two examinations.

There were two hypotheses in the current study. The first hypothesis was that there would be a high inter-marker correlation between pairs of markers in each module, with correlations comparable to that achieved in the other double-marked HKALE examination, the *Writing* paper. i.e., in the region of 0.75. The inter-marker correlation that emerged between pairs of markers for each module was in the moderate-to-strong region of 0.6, with lower correlations reported – generally in the moderate 0.4-0.5 level range – for inter-marker correlations on the question subparts. This hypothesis was, therefore, not proven.

Regarding the double marking of English public examination essay scripts, the system in place automatically invokes a third marker where there is a discrepancy of 5/24 marks. A similar system is planned for LS, where a third marking will be triggered at a discrepancy of 5/25 marks. The second hypothesis was that an overall discrepancy rate of just under 10% would emerge in the current LS study, comparable to the situation for English language. In the study, an overall discrepancy figure of 9.8% was achieved, and the hypothesis was accepted. Thus, while only

moderate correlations of 0.6 emerged in the current study, an equitable system would appear to be in place to compensate for such less-than-strong correlations.

The question remains whether 0.6 is high enough. Researchers such as Hatch and Lazaraton (ibid), suggest that correlations much below 0.8 indicate potential validity problems, with markers marking to different internal constructs. Indeed, if the lower correlations can be attributed to the single holistic marking criterion Content, this construct could well be considered under-specified. On this issue, Kuo (2007) argues strongly for the development of an analytic rubric for LS. Kuo (2007) asserts that by more succinctly defining 'critical thinking', 'analytic approach', 'cogent argumentation' etc, markers will be provided with more focus and direction, which will yield more reliable marking results (see also Weir, 1980). If this recommendation is acted upon, a follow-up study will therefore be required to verify whether analytic rubrics do increase marking reliability.

Double marking does offer tangible improvements in monitoring reliability and accounting for problematic marking. While there is currently only a moderate correlation between markers, the implementation of double marking as a system does enable the implementation of the additional monitoring process – the discrepancy rate third marker system – to operate as a safeguard for candidates. The exact manner in which double marking will be implemented in the HKDSE LS examination from 2012 onwards is not yet totally clear. In part, this is for three reasons. First, the format of the HKDSE will be different from that of the HKALE. Second, the marking scheme is more criterion-referenced, with marks for specific elements of content more explicitly detailed for different score bands. Third, the examination will be double marked through an onscreen marking system. It is likely, however, that the double marking procedures will generally mirror the system adopted for English language; that is, that where there is a discrepancy greater than 20% between any two markers' marks, the script will be sent to a third marker.

As will be appreciated, with the marking panel rising from its current 50 markers to as many as 700 when the candidature increases to 80,000 in 2012, the financial implications of double marking in the Hong Kong context are substantial. It is, nonetheless, resources which the HKEAA is prepared to allocate in order to convince stakeholders that concerns about different aspects of LS as a compulsory subject in the HKDSE are being attended to. In particular, it is important to convince the public that every effort is being made to resolve issues of the reliability of the marking before LS is examined with its full candidature in 2012.

As stated earlier, with the associated issues of reliability and discrepancies in marking, the implementation of double marking in a largely expanded public examination, along with the study of the incidence of discrepancy scripts, has implications and possible lessons for any examination authority that might be contemplating the introduction (or extension) of double marking in their public examination system.

Notes

1. The LS examination comprises six modules from which candidates sit two. Each module has one 2½-hour paper with three compulsory questions and one elective question to be chosen from the four available. Questions require an extended written response, with 20 marks allocated to Content and 5 marks to *Effective Communication* (logical argumentation, relevance of points made, effective use of language). The examination is single-marked, although this will change to double-marking in 2012.

2. In Hong Kong's schools, as the medium of instruction can be either English or Chinese, the examination has parallel English and Chinese versions of the question paper, and candidates may choose whether they wish to write their answers in English or Chinese for each of the two modules selected. In 2009, 93.8% of the 5,968 marked modules were answered in Chinese with 6.2% of modules answered in English (http://www.hkeaa.edu.hk/en/hkale/Exam_Report/).

Acknowledgement

I would like to thank the Hong Kong Examinations and Assessment Authority – and in particular Christina Lee, the General Manager for Assessment Development, and Lo Ka-yiu, the Senior Manager in charge of Liberal Studies – for support on the project regarding access to test takers' data.

References

Brooks, V. (1980). Improving the reliability of essay marking: a survey of the literature with particular reference to the English language composition, CSE Research Project Report , 5 (Leicester, University of Leicester).

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies* 52, 1, 29-46.

Chan, D.W. (2005). Liberalizing Liberal Studies in pre-university education in Hong Kong: Leadership development and beyond. *Educational Research Journal*, 20 (1), 1-14.

Chiu, C.S. & Mak, K.W. (2006). A critical review on curriculum development of Liberal Studies in Hong Kong. Hong Kong: Hong Kong Institute of Educational Research, Faculty of Education. (in Chinese).

Choi, C.C. and Lee, C. (2010). Developments of English language assessment in public examinations in Hong Kong. In Cheng, Liying and Curtis, Andy (Eds.). English Language Assessment and the Chinese Learner. pp. 60-76. Routledge: NY.

Coniam, D. (1991). Essay marking: a comparison of criterion-referenced and norm-referenced marking. *Institute of Language in Education Journal*, 7, 154-164.

Coniam, D. (2008). An investigation into the effect of raw scores in determining grades in a public examination of writing. *Japan Association for Language Teaching Journal*. 30 (1), 69-84.

Coniam, D. (2009a). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation*, 15(3), 243-263.

Coniam, D. (2009b). Validating onscreen marking in Hong Kong. Asia Pacific Education Review. Available at: http://www.springerlink.com/content/l6663m43jn501317/.

Coniam, D. & Yeung, A. (2010). Markers' perceptions regarding the onscreen marking of Liberal Studies in the Hong Kong public examination system. Asia Pacific Journal of Education, 30 (3), 249–271.

Cox, R. (1967). Examinations and Higher Education: Survey of the Literature. London, Society for Research into Higher Education.

Education Bureau. (2007). Secondary School Places Allocation System 2007/2009. Notes for Parents on Central Allocation. Available at: http://docs.google.com/gview?a=v&q=cache:4O-f-3ag_DwJ:www.edb.gov.hk/FileManager/EN/Content_1579/notesforparents_090421_eng.pdf.

Education Bureau (EDB). (2009). Fine-tuning the medium of instruction for secondary schools. LC Paper No. CB(2)623/08-09(01). Available at: http://www.edb.gov.hk/index.aspx?langno=1&nodeID=7128.

Harmon, M., Smith, T., Martin, M., Kelly, D., Beaton, A., Mullis, I., Gonzalez, E. & Orpwood, G. (1997). Performance assessment in IEA'S third international mathematics and science study (TIMSS). TIMSS International Study Center, Boston College: Chestnut Hill, MA. Available at: http://timss.bc.edu/timss1995i/TIMSSPDF/PAreport.pdf.

Hatch, E. & Lazaraton, A. (1991). The Research Manual. Heinle and Heinle: Boston, MA.

Hong Kong Examination and Assessment Authority (HKEAA) (2008). 2008 HKALE: AS Liberal Studies Examination Report and Question Papers (with marking scheme). Hong Kong: Hong Kong Examination and Assessment Authority.

Hong Kong Examination and Assessment Authority (HKEAA) (2009) HKALE Entries and Results Statistics over the Years. Retrieved from http://www.hkeaa.edu.hk/en/hkale/Exam_Report/Examination_Statistics/.

Ip, W. H. C. (2010). Promoting critical thinking: Discussing the capacity of issue-enquiry approach in Liberal Studies. Hong Kong Institute of Educational Research,: The Chinese University of Hong Kong, Hong Kong.

Kuo, S.W.A. (2007). Which rubric is more suitable for NSS Liberal Studies? Analytic or holistic? *Educational Research Journal*. 22(2), 179 – 199.

Lam, C.C. & Zhang, S. (2005). Liberal Studies: An illusive vision. *Hong Kong Teachers Centre Journal*, 4, 35-42.

Linton, O. (2004) An optimal estimator of true mark under double blind marking. Department of Economics, London School of Economics and Political Science, London, U Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: mathematics and English, British Education Research Journal , 22, 405–420.

Partington, J. (1994). Double-marking students' work. *Assessment and Evaluation in Higher Education* , 19 (1), 57–60.

Pilliner, A.E.G. (1969). Multiple marking: Wiseman or Cox? *British Journal of Educational Psychology* , 39 (3), 313–315.

Rong, M.C. (2005). The implementation and challenges of Liberal Studies in education reform in Hong Kong. *Hong Kong Teachers Centre Journal*, 4, 43-53.

Tsang, W.K. (2006). In search of the meanings for Liberal Studies in Hong Kong Senior Secondary Education. Hong Kong: Hong Kong Institute of Educational Research, Faculty of Education. (in Chinese)

Uebersax, J. (2009). Statistical methods for marker agreement. Available at: http://www.john-uebersax.com/stat/agree.htm

Vickers, E. (2005). In search of an identity: The politics of history as a school subject in Hong Kong, 1960s-2005. Hong Kong: Comparative Education Research Centre, The University of Hong Kong.

Vidal Rodeiro, C. L. (2007). Agreement between outcomes from different double marking models. Cambridge Assessment Research Matters, Issue 7, pp. 28-34. Available at: http://www.cambridgeassessment.org.uk/ca/digitalAssets/136145_Research_Matters_4_Jun_2007.pdf.

Weir, C.J. (1980). Communicative language testing. Prentice Hall Regents: NJ.

Author

David CONIAM, Professor,
Department of Curriculum and Instruction,
Faculty of Education,
The Chinese University of Hong Kong
[coniam@cuhk.edu.hk]

Received: 9.3.11; accepted 12.4.11; revised 26.4.11; further revised 14.5.11.